ACE-it: a tool for genome-wide integration of gene dosage and RNA expression data

van Wieringen, W.N.; Belien, J.A.M.; Vosse, S.; Achame, E.M.; Ylstra, B.

published in
Bioinformatics
2006

DOI (link to publisher)
10.1093/bioinformatics/btl269

document version
Publisher's PDF, also known as Version of record

BIOINFORMATICS APPLICATIONS NOTE Vol. 22 no. 15 2006, pages 1919–1920
doi:10.1093/bioinformatics/btl269

Genome analysis

ACE-it: a tool for genome-wide integration of gene dosage and RNA expression data

Wessel N. van Wieringen[1], Jeroen A. M. Belien[2], Sjoerd J. Vosse[1], Eskeatnaf M. Achame[2] and Bauke Ylstra[2,*]

[1]Department of Mathematics, Vrije Universiteit, De Boelelaan 1081, 1081 HV, Amsterdam, The Netherlands and [2]Department of Pathology, VU Medical Center, PO Box 7057, 1007 MB, Amsterdam, The Netherlands

Received on February 21, 2006; revised on April 24, 2006; accepted on May 19, 2006
Advance Access publication May 26, 2006
Associate Editor: Martin Bishop

ABSTRACT

Summary: We describe a tool, called ACE-it (Array CGH Expression integration tool). ACE-it links the chromosomal position of the gene dosage measured by array CGH to the genes measured by the expression array. ACE-it uses this link to statistically test whether gene dosage affects RNA expression.

Availability: ACE-it is freely available at http://ibivu.cs.vu.nl/programs/acewww/

Contact: b.ylstra@vumc.nl

Supplementary Information: Programs, the manual and supplementary information are available on the website.

Gene dosage, among other factors, affects gene expression in tumors (Albertson *et al*., 2000; Pollack *et al*., 2002), and can be measured on a genome-wide basis at high resolution by array comparative genomic hybridization (array CGH) (Pinkel and Albertson, 2005). Except for visualization implementations (Pollack *et al*., 2002; Autio *et al*., 2003), and calculation of correlations without formal inference (Nigro *et al*., 2005), sophisticated techniques for the integration of array CGH with expression array data are limited. A statistical tool for the detection of genes whose expression is affected by gene dosage within a series of samples is currently not available. ACE-it tests whether on a particular chromosomal location the RNA expression ratios are affected by the gene dosage. The relation with gene dosage is tested by ACE-it for each gene on the expression array. For this purpose gene dosage is divided into three categories: loss, normal or gain, defined by means of a user chosen cut-off for smoothed data (Olshen *et al*., 2004; Jong *et al*., 2004). Subsequently, the chromosomal positions are divided into two groups: positions at which the samples have either normal and gain, or normal and loss. The grouping disregards positions at which all the samples fall into one category, have a balanced distribution over the categories, or have no normals. The restriction into these two groups of chromosomal positions is biologically motivated by the assumption that for a given chromosomal location either oncogenes or tumor suppressor genes drive the chromosomal gain or loss (Albertson *et al*., 2000; Pinkel and Albertson, 2005). ACE-it allows a user-defined cut-off for contaminating samples within the groupings.

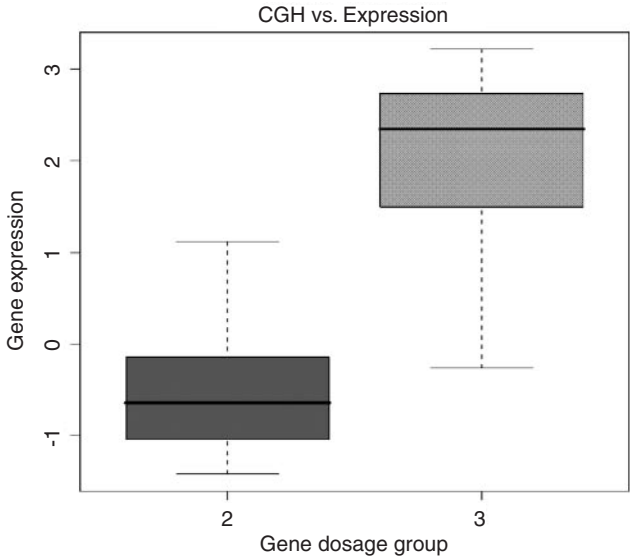

Fig. 1. ACE-it box plot of HER-2/neu at 17q12. *Y*-axis expression and *X*-axis gene dosage; normal, group 2; and gain, group 3. If this chromosomal position is gained, the expression is significantly higher.

Thus, normalized expression ratios for a particular gene are compared between the two categories within a group. ACE-it assumes that expression increases with increased gene dosage. This assumption leads to the null-hypothesis: (1) the median expression in samples with a gain is equal or smaller than that in samples with normal gene dosage; or, (2) the median expression in samples with a normal is equal or smaller than that in samples with a loss. We use the one-sided Wilcoxon's rank test to test this null-hypothesis, and apply the Benjamini–Hochberg's multiplicity correction to the resulting *P*-values (Benjamini and Hochberg, 1995). Genes whose adjusted *P*-value is smaller than the user-defined rejection level are considered differentially expressed between gene dosage levels.

*To whom correspondence should be addressed.

One compounding problem when linking gene dosage and RNA expression is that often the two arrays for expression and CGH are performed on different platforms (Ylstra *et al*., 2006). As a consequence, the arrays can have different elements spotted such that amount and chromosomal position of the ratios measured do not overlap. We solved this by imputing segmented ratios for each chromosomal position not covered by the arrayed elements, such that any given position on the chromosome gets assigned the smoothed array CGH value of the closest physical point of measurement. If expression and CGH array are performed on the same platform this feature may be disenabled.

ACE-it has been developed for the statistical software package R (http://www.r-project.org/) and a graphical user interface (GUI) for windows is provided. ACE-it was tested using array expression and array CGH datasets from various institutes and platforms, including a breast tumor series (Pollack *et al*., 2002). This yielded several genes whose expression is significantly affected by gene dosage, including HER-2/*neu* (c-*erb*B-2) which is amplified in ~20–30% of breast cancer cases (Fig. 1).

ACKNOWLEDGEMENTS

This work was in part supported by the Centre for Medical Systems Biology (CMSB), a centre of excellence approved by The Netherlands Genomics Initiative. Funding to pay the Open Access publication charges for this article was provided by the Centre for Medical Systems Biology (CMSB), a centre of excellence approved by The Netherlands Genomics Initiative.

Conflict of Interest: none declared.

REFERENCES

Albertson,D.G. *et al*. (2000) Quantitative mapping of amplicon structure by array CGH identifies CYP24 as a candidate oncogene. *Nat. Genet*., **25**, 144–146.

Autio,R. *et al*. (2003) CGH-Plotter: MATLAB toolbox for CGH-data analysis. *Bioinformatics*, **19**, 1714–1715.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate—a practical and powerful approach to multiple testing. *J. R Stat. Soc. B Methodol*., **57**, 289–300.

Jong,K. *et al*. (2004) Breakpoint identification and smoothing of array comparative genomic hybridization data. *Bioinformatics*, **20**, 3636–3637.

Nigro,J.M. *et al*. (2005) Integrated array-comparative genomic hybridization and expression array profiles identify clinically relevant molecular subtypes of glioblastoma. *Cancer Res*., **65**, 1678–1686.

Olshen,A.B. *et al*. (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, **5**, 557–572.

Pinkel,D. and Albertson,D.G. (2005) Array comparative genomic hybridization and its applications in cancer. *Nat. Genet*., **37**, S11–S17.

Pollack,J.R. *et al*. (2002) Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc. Natl Acad. Sci. USA*, **99**, 12963–12968.

Ylstra,B. *et al*. (2006) BAC to the future! or oligonucleotides: a perspective for micro array comparative genomic hybridization (array CGH). *Nucleic Acids Res*., **34**, 445–450.